Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

UNIVERSITY of
MASSACHUSETTS
AMHERST

Knowledge Discovery Lab
Department of Computer Science
Computer Science Building
(413) 545-3613
(413) 545-1249 (fax)

July 17, 2009

Defense Technincal Information Center
ATTN: BCS
8725 John J. Kingman Road, Suite 0944
Ft. Belvoir, VA 22060-0944

RE: HR0011-04-1-0013

Enclosed you will find the final technical report for the above referenced grant.

Please contact me if you have questions or require further documentation.

Sincerely,

Deb Bergeron
Grants Administrator
Knowledge Discovery Lab
David Jensen, Director
Bergeron@cs.umass.edu

The views, opinions, and/or findings contained in this article/presentation are those of the author/presenter and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.

Report Documentation Page	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE **17 JUL 2009**	2. REPORT TYPE **Final**	3. DATES COVERED **-**

4. TITLE AND SUBTITLE **Collective Inference with Learned and Engineered Knowledge**	5a. CONTRACT NUMBER **HR0011-04-1-0013**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Knowledge Discovery Lab Department of Computer Science Computer Science Building University of Massachusetts, Amherst**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S) **DARPA**
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**SAR**	**17**	

R & D Project Summary

Collective Inference with Learned and Engineered Knowledge

Tech/PI Contact
Prof. David Jensen
University of Massachusetts
140 Governors Drive
Amherst, MA 01003
Email: jensen@cs.umass.edu
Phone: (413) 545-9677
FAX: (413) 545-1249
Type of Business: Other Educational

Admin Contact
Ms. Carol Sprague
University of Massachusetts
Office of Grants and Contracts
Amherst, MA 01003
Email: sprague@resgs.umass.edu
Phone: (413) 545-0698
FAX: (413) 545-1202

Project URL:
http://kdl.cs.umass.edu/projects/

Table of Contents:

The views, opinions, and/or findings contained in this article/presentation are those of the author/presenter and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.

1 Project Information

1.1 Programmatic Information

ARPA Order Number:	Q894
Contract Number:	HR0011-04-1-0013
Performance Period:	February 12, 2004 – May 31, 2006
Project Title:	Collective Inference with Learned and Engineered Knowledge
Project URL:	http://kdl.cs.umass.edu/projects/
Performing Organization:	University of Massachusetts
Subcontractors:	None

Contacts:

Principal Investigator:

Prof. David Jensen
Computer Science Department, 140 Governors Drive
University of Massachusetts, Amherst, MA 01003 USA
Email: jensen@cs.umass.edu
Phone: (413) 545-9677, FAX: (413) 545-1249

Administrative:

Ms. Carol Sprague
Office of Grants and Contracts, Research Administration Bldg.
University of Massachusetts, Amherst, MA 01003 USA
Email: sprague@resgs.umass.edu
Phone: (413) 545-0698, FAX: (413) 545-1202

DARPA Program Mgr: Kendra Moore

1.2 Project Description

This section summarizes the corresponding text as presented in the proposal.

1.2.1 Research Objectives

1.2.1.1 Problem Description

A persistent goal of research in artificial intelligence has been to enable learning and reasoning with probabilistic models in complex domains. Much of this work has been directed toward systems that complement, rather than replace, human abilities and knowledge. Models that fuse *engineered knowledge* (knowledge from human sources) with *learned information* (information gained algorithmically) can take advantage of the strengths of both approaches, yielding more accurate predictions.

A particularly fruitful area for this research is improving our understanding of *emergent behavior*, specifically, how connectivity among individual units of a system affects global behavior. The Knowledge Discovery Laboratory (KDL) seeks to apply a growing understanding of emergent behavior to the design of learning and reasoning systems.

1.2.1.2 Research Goals

The main goal of this research was to apply network analysis to understand and improve the performance of relational dependency networks. Relational dependency networks (RDNs) are a new type of graphical model that exploit emergent behavior to improve both learning and inference. RDNs exhibit emergent behavior by using a set of individually learned conditional models (each estimating a probability distribution for a single attribute, given other attributes of the target object and related objects) to accurately estimate the full joint probability distribution over all attributes. RDNs also combine probabilistic and ontological reasoning, making it possible to easily integrate learned and engineered knowledge.

The specific goals for this project were:

1. Characterize the structure and dynamics of relational dependency network (RDN) models.
2. Reduce error of inferred probability distributions.
3. Increase efficiency of inference.
4. Produce domain-specific findings in at least two areas.
5. Produce open-source implementations of algorithms.
6. Develop new evaluation tools.

1.2.1.3 Expected Impact

The major benefit of the work is increasing the accuracy and efficiency of learned RDNs. Providing robust and scalable methods for learning and inference in RDNs would enable the development of modular and easily understood means for integrating learned and engineered knowledge in complex domains. Such techniques would have immediate applications in detecting financial fraud, understanding the structure of the Web, analyzing citation patterns in scientific papers and patent applications, and assessing organizational structures in government and business.

1.2.2 Technical Approach

1.2.2.1 Detailed Description of Technical Approach

Our approach centered on analyzing and understanding the characteristics of and relationships among the three different types of networks underlying RDNs: a *data network* representing the interrelated objects that serves as the input data for learning conditional models of statistical dependencies among attributes of those objects; a *model network* representing statistical dependencies among abstract variables that is constructed based on the learned conditional models; and an *inference network* representing statistical dependencies among specific random variables. Improving our understanding of these underlying networks for other models has resulted in improvements to those models; the same benefit is anticipated for learning and reasoning with RDNs.

We proposed extending our existing theoretical analysis of how the properties of data networks affected algorithms for learning conditional models from relational data. Some of our earliest work (e.g., Jensen 1999) showed how the structure of a data network affects feature selection in relational learning, and more recent work on autocorrelation (Jensen & Neville, 2002) and degree disparity (Jensen, Neville & Hay, 2003) has further added to our understanding of these effects. We have also shown that qualitative models of these influences allowed us to design and implement improved algorithms for learning conditional models (Neville, Jensen, Friedland, & Hay 2003). This area of inquiry promises similar benefits for new models such as RDNs.

We also proposed extending our relational algorithms to exploit ontological information during learning and to provide feedback to human authors of ontologies. Ontological information is commonly used in aggregating attribute values, a common technique for working with attributes from relational instances with variable structure. For example, a conditional relational model that estimates the probability a given paper will be published might aggregate attributes of related papers such as author count or page length. The ontology laid over the data network determines whether we consider all related papers together in such aggregations, or whether we distinguish between related papers by different authors and related papers by the same authors. Using different levels of an ontology establishes different equivalence classes and reinterprets the data for a relational learner. Course-grained levels of an ontology reduce the statistical variance associated with aggregated features, and fine-grained levels of an ontology reduce the statistical bias associated with aggregating potentially dissimilar objects. Either of these effects could allow construction of far more accurate conditional models, and a learning algorithm might search to find the right balance between these effects. In addition, such use of an ontology could indicate what levels of an ontology are most useful for learning, providing guidance to a human expert about where to elaborate an exiting ontology.

Because RDNs are assembled directly from conditional models, nearly anything that affects the structure of the conditional models will affect the structure of the model network represented by an RDN. Our previous research on RDNs focused on learning accurate conditional models; for this project we proposed a more systematic study of the structure of RDNs and the correlation between that structure and the characteristics of conditional learning algorithms. Specifically, we proposed applying standard network metrics and analysis algorithms to both the structure of RDNs and the strength of the probabilistic dependencies they encode. This analysis allows us to quantitatively characterize the structure of RDNs for different data sets and to

The views, opinions, and/or findings contained in this article/presentation are those of the author/presenter and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.

characterize the connection between the properties of the learning algorithm and the properties of the model network.

The accuracy of inference in RDNs depends critically on the emergent properties of the inference network, which is based on the relationships encoded in the data network and the probabilistic dependencies encoded in the model network. We proposed applying network metrics and analysis algorithms to a large number of different inference graphs to discern regularities in their structure and the strength of the probabilistic dependencies that they encode, allowing us to theoretically characterize how different structures in the model networks and data networks combine to produce structures in the inference network. We planned to use our analysis of these properties to build theoretical models and to design new algorithms for learning conditional models, RDN construction, and RDN inference.

1.2.2.2 Comparison with Current Technology

Relational dependency networks are relational models that encode joint probability distributions. These characteristics make them well-suited for integrating learned models with engineered ontologies and inference rules. Most other models lack one or more of these key characteristics.

Propositional learners (the largest class of current technologies) such as decision trees, discriminant analysis, linear regression, Bayesian networks and conventional dependency networks integrate poorly with common forms of engineered knowledge. Propositional representations cannot express the rich relational dependencies that are common in the first-order and higher-order representations used in the knowledge engineering community, and they cannot make direct use of ontologies. Additionally, the independence assumptions underlying propositional representations makes collective inference impossible, limiting their accuracy in realistic domains.

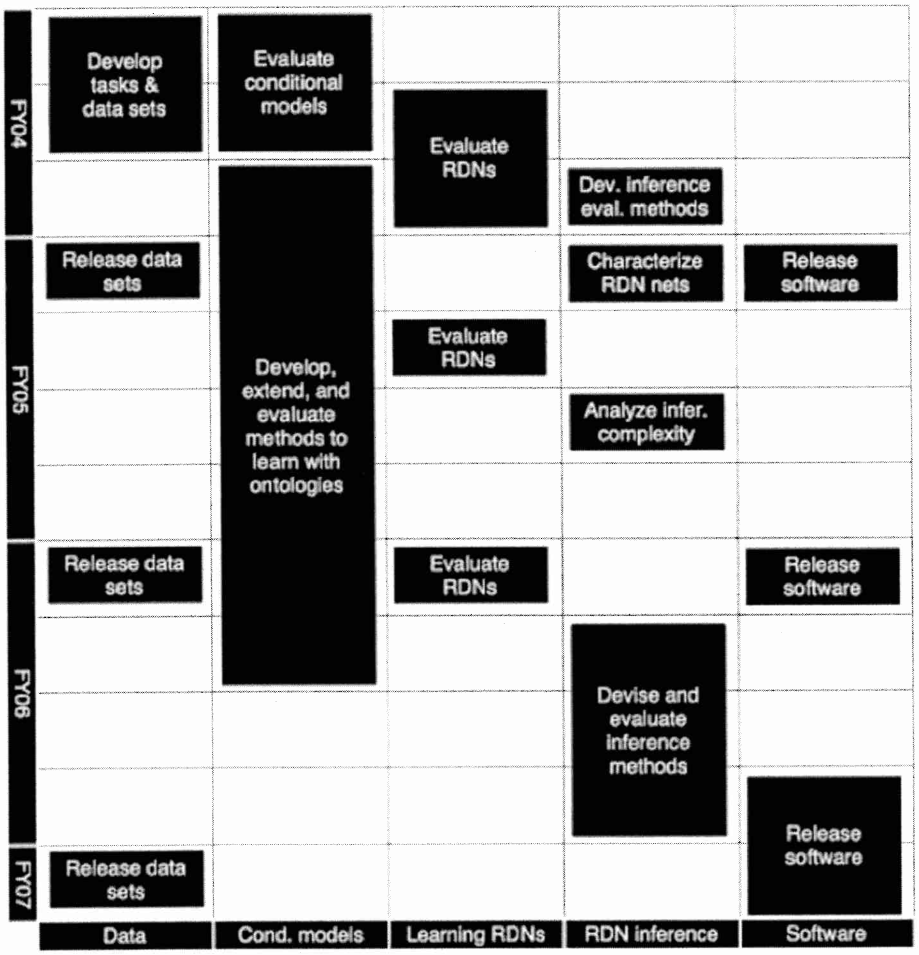
The remaining set of models can be classified as relational learning or inductive logic programming (ILP) models. Some of these models, such as traditional approaches to ILP, are strictly deterministic, eliminating the strength of probabilistic reasoning found in many learned models. Others, such as relational probability trees (Neville, Jensen, Friedland, & Hay 2003), 1BC2 (Lachiche & Flach 2003), and the relational neighbor classifier (Macskassy & Provost 2003), can only encode conditional probability distributions rather than joint distributions. Joint models compactly encode a larger number of probabilistic dependencies than conditional models and they allow multiple forms of reasoning, including both prediction and anomaly detection. Of the set of relational probabilistic models that encode joint distributions, two issues can severely limit their usefulness: Models may not encode coherent joint distributions (e.g., Bayesian logic programs, Kersting & De Raedt 2000), or they may be directed models, which cannot encode cyclic dependencies (Getoor et al. 2001).

Only two types of models meet all these criteria: relational Markov networks (Taskar et al. 2002) and RDNs. Understandability is a key requirement for any model intended to combine engineered and learned knowledge. Although both models can incorporate engineered knowledge, we have chosen to focus on RDNs because they are simple to describe and easier for domain experts to understand than relational Markov networks. Additionally, RDNs offer independent learning of components whereas relational Markov networks are not selective and require hand-coded features.

1.2.3 Schedule and Milestones

1.2.3.1 Schedule Graphic

The following graphic indicates the project milestones as presented in the proposal. Note that Q4 FY06 and Q1 FY2007 are beyond the scope of the funded project.



1.2.3.2 Detailed Individual Task Descriptions

The following tasks are from the proposal, with overall categories: (1) Data sets; (2) Learning and ontologies; (3) RDN construction; (4) RDN inference; and (5) Software. The percentage of that quarter's effort devoted to the effort in parentheses.

Q2 FY04

1.1 Data sets: Identify learning tasks (10%)
1.2 Data sets: Develop initial data sets (50%)
2.1 Learning and ontologies: Develop evaluation methods for learning conditional models (20%)
2.2 Learning and ontologies: Evaluate current learning of conditional models (20%)

Q3 FY04

1.2 Data sets: Develop initial data sets (continued) (20%)
2.2 Learning and ontologies: Evaluate current learning of conditional models (continued) (20%)
2.3 Learning and ontologies: Obtain ontologies for evaluation data sets (10%)
3.1 RDN construction: Develop evaluation methodologies for RDNs (20%)
3.2 RDN construction: Evaluate RDN learning with current conditional models (30%)

Q4 FY04

2.4 Learning and ontologies: Extend current learning algorithms to search ontology space (50%)
3.1 RDN construction: Develop evaluation methodologies for RDNs (continued) (10%)
3.2 RDN construction: Evaluate RDN learning with current conditional models (continued) (10%)
3.3 RDN construction: Create engineered conditional models (10%)
4.1 RDN inference: Develop evaluation methodologies for inference algorithms (20%)

Q1 FY05

1.3 Data sets: Release initial data sets (10%)
2.4 Learning and ontologies: Extend current learning algorithms to search ontology space (continued) (30%)
2.5 Learning and ontologies: Evaluate utility of fixed ontologies in learning (10%)
4.2 RDN inference: Characterize graphs of RDNs (20%)
5.1 Software: Release software (30%)

Q2 FY05

2.6 Learning and ontologies: Devise and implement methods to assist ontology construction (70%)
3.4 RDN construction: Evaluate RDN learning with fixed ontologies (30%)

Q3 FY05

2.7 Learning and ontologies: Develop methods for co-learning conditional models & ontologies (60%)
4.3 RDN inference: Develop theoretical analysis of inference complexity (40%)

Q4 FY05

2.7 Learning and ontologies: Develop methods for co-learning conditional models & ontologies (continued) (60%)
2.8 Learning and ontologies: Evaluate utility of co-learning conditional models & ontologies (40%)

Q1 FY06

1.4 Data sets: Revise data sets (20%)
2.7 Learning and ontologies: Develop methods for co-learning conditional models & ontologies (continued) (20%)
2.9 Learning and ontologies: Improve efficiency of co-learning methods (30%)
3.5 RDN construction: Evaluate RDN learning with learned ontologies (10%)
5.2 Software: Release software (20%)

Q2 FY06

2.9 Learning and ontologies: Improve efficiency of co-learning methods (continued) (50%)
4.4 RDN inference: Devise methods to decompose rolled-out model network (50%)

Q3 FY06

4.4 RDN inference: Devise methods to decompose rolled-out model network (continued) (60%)
4.5 RDN inference: Evaluate efficiency and accuracy for simplified inference (40%)

Q4 FY06 (Project ended before this quarter)

4.5 RDN inference: Evaluate efficiency and accuracy for simplified inference (continued) (50%)
5.3 Software: Release software (50%)

Q1 FY07 (Project ended before this quarter)

1.5 Data sets: Release revised data sets (50%)
5.3 Software: Release software (continued) (50%)

1.2.4 Deliverables Description

The proposed statement of work listed the following deliverables. A brief description of actual performance is given in italics. For specific details, see section 3.1.5.

- **Software releases** — As specified in the schedule, we will release new versions of Proximity approximately every 12 months. Additional releases are likely, as new incremental capabilities are produced.

  KDL produced six major releases (at approximately six-month intervals) of Proximity during the contract period.

- **Data sets** — As specified in the schedule, we will release new benchmark data sets to aid comparative studies and replication of our technical work.

  KDL released four benchmark data sets during the contract period.

The views, opinions, and/or findings contained in this article/presentation are those of the author/presenter and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.

- **Technical papers** — Each new major technical finding and algorithm will be reported in technical papers submitted to workshops, conferences, and scholarly journals. All such papers will be made available on the PI's website, except where forbidden by copyright restrictions.

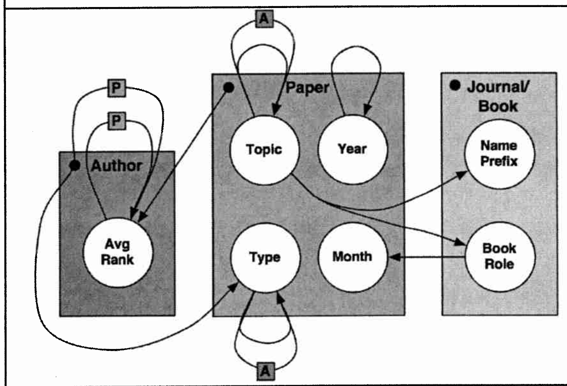
  KDL published twelve technical papers in journals and workshop and conference proceedings during the contract period.

Software and data sets were released open-source, and papers and other written products are available under ordinary copyright agreements.

Technology Transition and Technology Transfer, Targets and Plans

KDL's primary venue for technology transfer is through releases of its Proximity software. As well as providing periodic open-source releases, KDL works with government and commercial organizations to apply our techniques to real analysis tasks in complex domains.

1.2.5 Quad Chart

<table>
<tr><th colspan="2">Collective Inference with Learned and Engineered Knowledge</th></tr>
<tr><td>
</td><td><b>NOVEL IDEAS</b>
• Relational Dependency Networks (RDNs) express <i>statistical dependence among characteristics of related entities</i> (e.g., authors and the papers they write).
• RDNs can be <i>learned far more efficiently</i> than other similarly expressive models.
• RDNs <i>allow domain experts to express prior knowledge</i> in natural and highly expressive ways.</td></tr>
<tr><td><b>IMPACT:</b>
• Decrease by an order of magnitude the time required to learn an accurate relational model of a domain over alternative methods (e.g., relational Markov networks (RMNs)).
• Learn models of equivalent or higher accuracy than alternative relational models (e.g., RMNs)
• Learn models with 50% reduction in error over models using conditional inference (e.g., Relational Probability Trees).</td><td><b>SCHEDULE:</b>
(see section 1.2.3)</td></tr>
</table>

The views, opinions, and/or findings contained in this article/presentation are those of the author/presenter and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.

2 Funding Report

2.1 Funding Obligated this Period

$1,000,000

2.2 Funding Obligated to Date

$1,000,000

2.3 Incurred Expenses this Period

$1,000,000

2.4 Incurred Expenses to Date

$1,000,000

2.5 Invoices this Period

$0

2.6 Invoices to Date

$1,000,000

3 Technical Report

3.1 Project Progress

Summary

During the course of the project, we conducted extensive tests of relational dependency networks, publishing the results in a journal article (Neville & Jensen 2007). We developed a new type of probabilistic model for relational data, called a latent group model, and extended another of our existing models, relational probability trees, to handle probabilistic dependencies that involve time. We studied how to model the probability of link occurrence based on the pattern of surrounding links; such models are an important complement to probabilistic models of the values of attributes given relational structure. We applied our work on relational learning to several real-world challenge problems, including predicting fraud among stock brokers, reasoning about robotic movements, and constructing peer-to-peer networks.

Details for each of these accomplishments are provided below.

Personnel actions

KDL graduated two students (one Ph.D. and one M.S.), saw a post-doc researcher move on to a tenure-track faculty position, and accepted two new graduate students.

- Amy McGovern, a postdoc in KDL since Fall 2002, joined the University of Oklahoma as an Assistant Professor of Computer Science for the fall term, 2004.
- Jennifer Neville, KDL's first doctoral student, successfully defended her dissertation proposal on October 21, 2005. She completed her dissertation on August 1, 2006 and accepted a tenure-track faculty position at Purdue University. Her dissertation, *Statistical Models and Analysis Techniques for Learning in Relational Data*, was nominated for an ACM Doctoral Dissertation Award.
- Brian Gallagher completed his M.S. in Computer Science in January, 2006. He currently works on developing algorithms and analysis techniques for complex networks at the Center for Applied Scientific Computing at Lawrence Livermore National Laboratory.
- Two new graduate students, Brian Taylor and Marc Maier, joined KDL in September, 2005.

3.1.1 Progress Against Planned Objectives

KDL has met the proposed objectives (listed in section 1.2.1.2) as noted below:

Characterize the structure and dynamics of relational dependency network (RDN) models.

We conducted an extensive experimental evaluation of relational dependency networks. This analysis looks at RDNs in the context of relational Bayes networks and relational Markov networks and demonstrates the relative strengths of RDN models, namely, the ability to represent cyclic dependencies, simple methods for parameter estimation, and efficient structure learning. The evaluation also included learning RDNs for a number of real-world datasets and evaluating the models in a classification context, where only a single attribute is unobserved. In addition, we used synthetic data to explore model performance under various relational data characteristics.

Our results showed that, except in rare cases, the performance of RDNs approaches the performance that would be possible if all the class labels of related instances were known. However, our analysis indicates that the amount of seed information may interact with the level of autocorrelation and local model characteristics to impact performance. Future work will attempt to quantify these effects more formally. The full details of this evaluation are presented in a *Journal of Machine Learning Research* article (Neville & Jensen 2007).

Reduce error of inferred probability distributions.

We implemented a number of small improvements to the algorithms for learning conditional probability distributions. These improvements, in turn, affect the accuracy of the resulting RDNs. This objective was de-emphasized during the project as it became clear that the unique research issues of RDNs primarily concerned collective inference rather than improved conditional models.

Increase efficiency of inference.

We studied the convergence properties of RDNs, and concluded that RDN inference is relatively efficient despite the use of Markov-chain Monte Carlo approaches. RDN inference networks tend to converge quickly to accurate probability distributions, thus we focused more on evaluation and characterization of the circumstances in which collective inference performs effectively.

Produce domain-specific findings in at least two areas.

We applied our work on relational learning to several real-world challenge problems, including predicting fraud among stock brokers, reasoning about robotic movements, and constructing peer-to-peer networks.

- The first project was conducted jointly with the National Association of Securities Dealers (NASD), who provided extensive access to data and expertise of their analysts (Neville at al. 2005). We conducted an extensive analysis of data from NASD about stock fraud, applying several different analysis techniques and combining the results into a statistical model that predicts the probability of fraud for individual brokers. In addition, we conducted an extensive evaluation of the results on new data, using four person-weeks of time from NASD professional examiners to evaluate the utility of the results and comparing them to current NASD screening rules. Model predictions were found to correlate highly with the subjective evaluations of experienced NASD examiners. Furthermore, in all performance measures, our models performed as well as or better than the handcrafted rules that are currently in use at NASD.
- The second project focused on applying relational dependency networks to predicting the outcomes of movements of a robotic torso and was conducted jointly with the UMass Laboratory for Perceptual Robotics (Hart, Grupen & Jensen 2005) . The resulting predictions were consistent with the training data and yielded a policy that allowed picking up two differently shaped objects correctly.
- The third project focused on construction of social networks to improve peer-to-peer networking and was conducted jointly with the UMass Privacy, Internetworking, and Mobile Systems Laboratory (Fast, Jensen & Levine 2005). We studied methods of constructing peer-to-peer networks based on the preferences of users. The approach uses a model of user preference identified by latent-variable clustering with hierarchical Dirichlet processes (HDPs) to identify users who are likely to trade files in the future. Our simulations and empirical studies show that the clusters of songs created by HDPs effectively model user behavior and can be used to create desirable network overlays that outperform alternative approaches.

Produce open-source implementations of algorithms.

KDL released several versions of Proximity, an open-source environment for relational knowledge discovery. Proximity includes implementations of our models including the relational Bayesian classifier, relational probability trees, and relational dependency networks. See section 3.1.5, for additional information on KDL's Proximity releases.

Develop new evaluation tools.

Because both the learning and inference processes can introduce errors, relational learning algorithms pose new challenges for error analysis, necessitating a new framework for analyzing and decomposing these errors. We developed a new bias/variance framework that decomposes loss into errors due to both the learning and inference process. We evaluated performance of three relational models on synthetic data and used the framework to understand the reasons for poor model performance. With this understanding, we proposed a number of directions to explore to improve model performance. The full details of this framework are presented in "Bias/variance analysis for network data" (Neville & Jensen 2006).

3.1.2 Technical Accomplishments

In addition to the accomplishments against planned objectives noted above, we produced the following unplanned accomplishments:

Latent Group Models

We invented and implemented a new type of probabilistic model for relational data, called a *latent group model*. Latent group models represent autocorrelation dependencies by hypothesizing a hidden group entity. This greatly simplifies both parameter learning and inference, resulting in a much more tractable probabilistic model than conventional probabilistic relational models or relational Markov networks. We conducted an experimental evaluation of latent group models to determine their accuracy and performance characteristics.

Link Prediction

We studied how to model the probability of link occurrence based on the pattern of surrounding links. Such models are an important complement to probabilistic models of the values of attributes given relational structure. They are also useful in their own right for tasks such as predicting whether two individuals will collaborate in the future, whether someone willl purchase a product, or whether someone will be interested in a given document.

Temporal Extensions to Relational Models

We examined and experimented with methods of extending our existing models of relational data to handle probabilistic dependencies that involve time. The result is a temporal extension to our relational probability trees. Evaluation of our new algorithms for learning temporal features for relational probability trees revealed a strong potential for overfitting and new type of errors specific to temporal data.

3.1.3 Improvements to Prototype

Our software environment for relational knowledge discovery, Proximity, was completely re-written to convert to a new underlying database structure, MonetDB, a fast, open-source vertical database. The switch to MonetDB resulted in orders of magnitude speed improvements for the kinds of operations needed by relational knowledge discovery compared to implementations hosted on SQL databases.

We have also added numerous new features and capabilities to Proximity including:

- an implementation of relational dependency networks
- support for temporal features in relational probability trees
- cleanup of model code, resulting in improved ease of use and significant increases in speed
- extensions to Proximity's implementation of the graphical query language QGraph
- social networking analysis tools
- synthetic relational data generation capabilities
- new aggregation functions
- additional capabilities for and finer control of data import and export functionality

Usability improvements included a completely new graphical user interface that eliminated some earlier cumbersome requirements and provided new tools for exploring, analyzing, modeling, and visualizing data. We have also written a new graphical editor for creating QGraph queries, added an interactive interpreter for executing short scripts, and created professionally written documentation for both Proximity and QGraph.

3.1.4 Significant Changes to Technical Approach

The focus of the project shifted somewhat from making improvements in the mid-range technologies for RDNs to the "edges" — both the theoretical (making fundamental theoretical advances in the understanding of why RDNs perform so effectively despite their very simple approaches to learning and inference) and the applied (evaluating the performance of RDNs in practical inference tasks). This shift resulted from our early evaluation results that showed that RDNs were surprisingly effective in their current form.

3.1.5 Deliverables

We proposed three types of deliverables: software releases, benchmark data sets, and technical publications. These deliverables have been met as described below. (See section 1.2.4 for a description of the proposed deliverables.)

Software releases

KDL produced major releases of its open source software, Proximity, twice a year during the course of the contract.

- Proximity 3.0 (April 15, 2004) — First open-source release of Proximity after a complete rewrite of the code, featuring a new underlying database architecture (an open-source vertical database called MonetDB), a scripting interface for programmers and users, and a professionally written tutorial

- Proximity 3.1 (September 15, 2004) — Added visualizer for relational probability trees, interactive interpreter for our Python-based scripting language, and the ability to import and export XML-formatted data
- Proximity 4.0 (March 2, 2005) — First release of our implementation of relational dependency networks including a viewer for RDNs, a new QGraph editor, a generator of relational data useful for experimentation, a substantial redesign of the API, and new documentation for QGraph
- Proximity 4.1 (December 13, 2005) —Substantially improved GUI to ease experimentation, an updated API to ease experimental analysis, new methods for measuring performance of inference procedures for RDNs, and updated QGraph guide
- Proximity 4.2 (April 24, 2006) — Added methods to create temporal features in our relational probability tree models and an updated API to ease experimental analysis

Data sets

KDL released the following data sets:

- HEP-Th — Data on papers in high-energy physics derived from the abstract and citation files provided for the 2003 KDD Cup competition. The original datasets are from arXiv, an electronic archive of research papers physics and selected other sciences, and the SLAC SPIRES-HEP database, a comprehensive catalog of high-energy particle physics literature compiled by the Stanford Linear Accelerator Center. The dataset contains over 42,000 objects, over 500,000 links, 39 object attributes, and 15 link attributes.
- Can-o-sleep — Records of all the mp3 files shared by and transferred between users during an 81-day period in the spring of 2003. The dataset contains over 500,000 objects, over 6 million links, 14 object attributes, and 6 link attributes.
- DBLP — Information on computer science publications listed in the DBLP Computer Science Bibliography derived from a snapshot of the bibliography as of April 12, 2006. The dataset contains over 1,200,000 objects, over 2,480,000 links, 12 object attributes, and 6 link attributes.
- Mobile Social Networks — Data taken from a series of experiments in wireless mobile connections undertaken by the Privacy, Internetworking, Security, and Mobile Systems Laboratory at the University of Massachusetts Amherst. The dataset contains 27 objects, over 180,000 links, 1 object attribute, and 2 link attributes.

Technical papers

KDL published twelve technical papers that report the results of research performed as part of this project. The papers are listed in section 3.1.7.

3.1.6 Technology Transition and Transfer

3.1.6.1 Technology Transition and Transfer Description

All transitioned technologies were implemented with versions of Proximity, our open-source environment for relational knowledge discovery. Details of the capabilities of Proximity can be found in the Proximity Tutorial, QGraph Guide, Cookbook, and Javadoc.[1]

3.1.6.2 Technology Transition and Transfer List

Our Proximity software was downloaded more than 10,500 times during the contract period, though it is not possible to track all the organizations that may be evaluating or using Proximity. That said, 46 of the downloads came from .mil domains and 21 came from .gov domains. From support requests, we know that several government laboratories, university research groups, and large government contractors have evaluated or are actively using Proximity.

3.1.7 Publications

Fast, A., D. Jensen, and B.N. Levine (2005). Creating social networks to improve peer-to-peer networking. *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

[1] Available at: http://kdl.cs.umass.edu/proximity/documentation.html

Fast, A., and D. Jensen (2006). The NFL coaching network: Analysis of the social network among professional football coaches. *AAAI Fall Symposium on Capturing and Using Patterns for Evidence Detection*.

Hart, S., R. Grupen, and D. Jensen (2005). A relational representation for procedural task knowledge. *Proceedings of the 20th National Conference on Artificial Intelligence*. Pittsburgh, Pennsylvania.

Neville, J., Ö. Simsek, D. Jensen, J. Komoroske, K. Palmer, and H. Goldberg (2005). Using relational knowledge discovery to prevent securities fraud. *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Neville, J., and D. Jensen (2005a). Leveraging relational autocorrelation with latent group models. *Dagstuhl Seminar 05051: Probabilistic, Logical and Relational Learning—Towards a Synthesis*.

Neville, J., and D. Jensen (2005b). Leveraging relational autocorrelation with latent group models. *Proceedings of the 4th Multi-Relational Data Mining Workshop, 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Neville, J., and D. Jensen (2005c). Leveraging relational autocorrelation with latent group models. *Proceedings of the 5th IEEE International Conference on Data Mining*.

Neville, J., and D. Jensen (2006). Bias/variance analysis for network data. *Proceedings of the Workshop on Statistical Relational Learning, 23rd International Conference on Machine Learning*.

Neville, J., and D. Jensen (2007). Relational Dependency Networks. *Journal of Machine Learning Research*. 8: 653-692.

Neville, J., and D. Jensen (2007). Relational dependency networks. *Introduction to Statistical Relational Learning*, L. Getoor and B. Taskar, editors.

Rattigan, M., and D. Jensen (2005a). The case for anomalous link detection. *Proceedings of the 4th Multi-Relational Data Mining Workshop, 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Rattigan, M., and D. Jensen (2005b). The case for anomalous link discovery. *ACM SIGKDD Explorations*, vol. 7 issue 2, December, 2005.

3.1.8 Meetings and Presentations

April 12-13, 2004 – The PI attended a workshop on new potential DARPA programs in machine learning in Philadelphia, PA. He presented KDL's current work on relational learning and proposed several potential challenge problems.

June 10, 2004 – The PI spoke at a workshop on social network analysis held at Yahoo Labs in Pasadena, CA. His talk presented results of early work on relational dependency networks.

September 22-23, 2004 – The PI presented a talk on relational learning methods at the DHS Data Sciences Workshop in Washington, DC.

October 3, 2004 – The PI presented a talk at the "Turning Information into Knowledge" workshop sponsored by the International Atomic Energy Agency in New York, NY.

January 31-February 4, 2005 – The PI and a graduate student (Jennifer Neville) presented two talks at the Dagstuhl seminar on Probabilistic, Logical and Relational Learning. Neville's talk was on "Leveraging relational autocorrelation with latent group models" and Jensen's talk examined the question "Does Accurate Statistical Inference Require Joint Models of Attributes and Relations?" The invitation-only seminar was a major week-long international meeting of researchers in relational learning held in southwestern Germany.

March 7-8, 2005 – The PI participated in a workshop convened by the U.S Treasury and NSF entitled "Resilient Financial Information Systems" in Washington, DC.

May 31-June 1, 2005 – The PI gave an invited talk at the Pacific Northwest National Laboratories in Pasco, Washington. The talk discussed learning probabilistic models and implications for data visualization.

The views, opinions, and/or findings contained in this article/presentation are those of the author/presenter and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.

Summer, 2005 – Presented papers at major conferences and workshops — Students gave six presentations of accepted papers at major conferences and workshops (Fast, Jensen, and Levine 2005; Hart, Grupen, and Jensen 2005; Neville et al. 2005; Neville and Jensen 2005b; Neville and Jensen 2005c; Rattigan and Jensen 2005a).

July 9-10, 2005 – Participated in AAAI Doctoral Consortium — KDL students Jennifer Neville and Özgür Simsek presented their work at the Tenth AAAI/SIGART Doctoral Consortium in Pittsburgh, Pennsylvania during the Twentieth National Conference on Artificial Intelligence. Their presentations were "Structure Learning for Statistical Relational Models" and "Towards Competence in Autonomous Agents". The AAAI and ACM/SIGART Doctoral Consortium provides an opportunity for a group of Ph.D. students to discuss and explore their research interests and career objectives with a panel of established researchers in artificial intelligence.

September 26-27, 2005 – Attended National Academy of Sciences meeting — The PI gave an invited talk at a National Academy of Sciences meeting on Statistics on Networks in Washington, DC.

October 11-12, 2005 – The PI participated in a DARPA workshop on Behavioral Economics and Networks in Philadelphia, PA.

October 18-19, 2005 – The PI participated in a DARPA DSO workshop on Virtual Worlds in Washington, DC.

October 20, 2005 – The PI gave an invited talk at the Computer Science Department at Tufts University.

3.1.9 Issues or Concerns

(None)

3.2 Project Plans

The contract is completed.

4 References

Fast, A., D. Jensen, and B.N. Levine (2005). Creating social networks to improve peer-to-peer networking. *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Getoor, L., N. Friedman, D. Koller and A. Pfeffer (2001). Learning probabilistic relational models. In Dzeroski and Lavrac (Eds.), *Relational Data Mining*, Springer-Verlag.

Hart, S., R. Grupen, and D. Jensen (2005). A relational representation for procedural task knowledge. *Proceedings of the 20th National Conference on Artificial Intelligence.* Pittsburgh, Pennsylvania.

Jensen, D. (1999). Statistical challenges to inductive inference in linked data. In *Papers of the 7th International Workshop on Artificial Intelligence and Statistics.*

Jensen, D., and J. Neville (2002). Linkage and autocorrelation cause feature selection bias. In *Proceedings of the 19th International Conference on Machine Learning.*

Jensen, D., J. Neville, and M. Hay (2003). Avoiding bias when aggregating relational data with degree disparity. In *Proceedings of the 20th International Conference on Machine Learning.*

Kersting, K. and L. De Raedt (2000). Bayesian logic programs. In J. Cussens and A. Frisch (Eds.), *Work-in-Progress Reports of the 10th International Conference on Inductive Logic Programming.*

Lachiche, N. and P. Flach (2003). 1BC2 : A true first-order Bayesian classifier. In S. Matwin and C. Sammut (Eds.), *Proceedings of the 12th International Conference on Inductive Logic Programming. Lecture Notes in Artificial Intelligence 2583*, Springer-Verlag. pp. 133-148.

Macskassy, S. and F. Provost (2003). A simple relational classifier. In Dzeroski, De Raedt, and Wrobel (Eds.). *Proceedings of the 2nd Workshop on Multi-Relational Data Mining.*

Neville, J., and D. Jensen (2006). Bias/variance analysis for network data. *Proceedings of the Workshop on Statistical Relational Learning, 23rd International Conference on Machine Learning.*

Neville, J., and D. Jensen (2007). Relational dependency networks. *Journal of Machine Learning Research.*

Neville, J., Ö. Simsek, D. Jensen, J. Komoroske, K. Palmer, and H. Goldberg (2005). Using relational knowledge discovery to prevent securities fraud. *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Neville, J., D. Jensen, L. Friedland and M. Hay (2003). Learning relational probability trees. In *Proceedings of the 9th International Conference on Knowledge Discovery & Data Mining.*

Taskar, B., P. Abeel and D. Koller (2002). Discriminative probabilistic models for relational data. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence.*